Chapter 23
Applications: simple models and difficult theorems

Nelly Litvak

Abstract In this short article I will discuss three papers written by Willem van Zwet with three different co-authors: Mathisca de Gunst, Marta Fiocco, and myself. Each of the papers focuses on one particular application: growth of the number of biological cells [3], spreading of an infection [7], and the optimal travel time in warehousing carousel systems [8].

23.1 Introduction

In this short article I will discuss three papers written by Willem van Zwet with three different co-authors: Mathisca de Gunst, Marta Fiocco, and myself. Each of the papers focuses on one particular application: growth of the number of biological cells [3], spreading of an infection [7], and the optimal travel time in warehousing carousel systems [8]. To my opinion, each of these papers displays the attitude that I personally value a lot in mathematics. An application is the strong starting point for each of the papers. Further, the model is simple and transparent. Yet, the analysis involves advanced mathematics and brings to the results that not only give new insights into the applications but also are of a pure mathematical interest. The present volume contains [7] and [8], and the follow-up paper [4] of [3] which I will also briefly discuss.

The papers are written in a clear language and do not try to look more fancy than they are. In fact, I remember Willem laughing at my attempts to make the paper more general by replacing 1/2 with $b \in (0,1)$: 'What have you done? Please, bring the $1/2$ back! It is more natural and makes the whole thing much easier to read'. And on my sceptical remark about the number of people who are actually going to

Nelly Litvak
Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
e-mail: n.litvak@ewi.utwente.nl

S. van de Geer and M. Wegkamp (eds.), *Selected Works of Willem van Zwet*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4614-1314-1_23, © Springer Science+Business Media, LLC 2012

read this text he smiled again: 'Well, you have to assume people will read it.' Now, assuming that people will read the introduction to this chapter, I will try, to the best of my own understanding, to describe the essence of the models and the results for each of the papers, what I think was difficult and why it worked. I will try to stick to common sense and intuition, so please forgive me if I am not very precise and go ahead, read the papers for correct formulations and exact results.

23.2 A non-Markovian model for cell population growth

Biostatistics is an extremely important topic, popularity of which has grown hugely in the last years. The paper [3] describes a model for a cell population growth. Initially, we have n plant cells transferred to a medium of a known composition at time $t = 0$. The cells can divide, and we are interested in the number $N_n(t)$ of cells at time $t > 0$. Specifically, we want to obtain a law of large numbers and a central limit theorem for the process $N_n(t)$ as n grows large. The motivation for this problem formulation is that in reality the number of cells is quite large.

The division happens as follows. From the medium, the cells receive a stimulus at a random time, and after that it takes a cell exactly c time units before it divides. The time it takes to receive a stimulus depends on the concentration of a substrate (sugar) in the medium. Clearly, with time, the substrate is being used up and thus it takes longer before a cell receives a stimulus. As described so far, the model already contains two non-trivial features. First, the rate at which the cells receive a stimulus is variable (non-increasing). Second, the cells' 'pregnancy' of length c obviously makes the process $N_n(t)$ non-Markovian. There is also a third interesting feature of the model, namely, the authors distinguish between A-cells and B-cells, where only A-cells are able to divide. As a result of a division, two cells are produced, each of which can be an A-cell with a probability that depends on the concentration of a hormone in the medium. Again, with time the hormone is being used up and thus the probability of producing an A-cell is decreasing.

Altogether, the model description is not hard and very natural but each of the model assumptions brings essential new features in the analysis. Then, what makes this model solvable? One helping feature is the 'boundedness' of the process. First of all, at most two cells can be born at each division. This makes the number of born A-cells bounded, and we can apply the inequalities of the type presented in Lemma 4.2, which resembles the Azuma's inequality for martingales (see e.g. [13, p. 307]). Second, the authors assume that the amount of the substrate and the hormone is proportional to the original number of cells. This is a natural scaling, which ensures that, on average, each cell can potentially receive a certain fixed amount of both ingredients. For each cell, this makes the whole process bounded. Therefore, intuitively, it is clear that after a random finite time T_n no division will happen for one of the two possible reasons: either the substrate is finished and thus no cell can receive a stimulus, or the hormone is finished and thus no more A-cell is born. Moreover, the total amount of cells remains of the order n at any time, which ensures that

the usual scalings for the large deviation result (Theorem 4.2) and the central limit theorem (Theorem 5.1) work in this setting. Another feature that makes the model tractable is that, despite the process being non-Markovian, the time it takes a new cell to obtain a stimulus is exponential, which allows to talk about the rates and use the bounds developed for Markov processes (e.g. Lemma 4.1).

The large deviation result established in Theorem 4.2 implies that that $N_n(t)/n$ converges to a function $X(t)$ in probability, uniformly in t at exponential rate when n grows to infinity. Here the function $X(t)$ is the averaged integrated intensity of the process. To obtain $X(t)$, the authors need to make several steps of conditioning and averaging, where the first important step is the conditioning on the number of A-cells produced at each division. Obviously, in this model, the intensity at time t depends on the aggregated intensity before time t because this aggregated intensity defines how much substrate and hormone has been used before t. Hence, it is natural that $X(t)$ is defined as a solution of an integral equation. Technically, the uniform convergence result is very difficult and requires a lot of preliminary work. Totally different argument is used to prove the convergence for a bounded t (Theorem 4.1) and for $t \to \infty$ (Lemma 4.6). Finally, the proof of the main theorem combines all the preliminary results plus uses a very elegant argument to control the deviation of the integrated intensity process from $X(t)$.

The central limit theorem in Section 5 describes in detail the convergence of the process $V_n(t) = n^{1/2}(N_n(t)/n - X(t))$ to its limit $V(t)$ in distribution, where the convergence is in the sense of the Skorohod metric. The process $V(t)$ involves two independent Wiener processes: one of them, W_0, is responsible for the random deviation of $N_n(t)$ from the integrated intensity process, and another one, W_1, reflects the randomness due to a random number of A-cells produced at each division. Clearly, $V(t)$ is again a solution of an integral equation that involves both W_0 and W_1 in a non-trivial way. The form of $V(t)$ and its covariance structure are really complicated, and, as noticed by the authors, 'almost impossible to guess without going into the special structure of the underlying process...'.

Last section contains numerical examples, which show that the scaling results are in good agreement with experimental data.

The above summarises paper [3], which is the first part of the analysis of the non-markovian model of the population growth. This volume contains the second part of this work, paper [4], where the duration of the growth is analysed. Here the authors obtain a remarkable discontinuity result. It turns out that with a certain balance between the initial amount of hormone and substrate the number of divisions and the duration of the process is much larger than for other values of the parameters. Another example of surprising properties of this deep interesting model.

23.3 Parameter estimation for the supercritical contact process

The paper studies a contact process on a d-dimensional grid. The model description is typical for processes of this sort. Each site in $\mathbb{Z}^d$ is either infected or healthy.

A healthy site gets infected at rate λ by any of its infected neighbors. An infected site becomes healthy at rate 1. The process is supercritical, that is, with positive probability, an infection started by one infected site, will last forever. This is ensured by the inequality $\lambda > \lambda_d$, where λ_d is a critical value.

The goal of the paper is to estimate the parameter λ. Intuitively, it is not hard to imagine what the estimator should be. The authors follow a most natural path. Provided that the process started by a single infected site in 0 and survives forever, we take some set D where a stationary regime has been established. Then the estimator for λ at time t is simply

$$\hat{\lambda} = \frac{\#\text{ infected sites in } D \text{ at } t}{\#\text{ sites in } D \text{ that are healthy but have infected neighbors}}. \tag{23.1}$$

The fraction above is a result of a balance equation in stationarity: the denominator multiplied by λ is the rate at which new sites get infected, and the nominator (multiplied by one) is the rate at which infected sites get healthy. In stationarity, both rates should be equal.

The description of the proposed estimator will be complete once we decide on how to choose D, and this is where the main difficulty lies because each of the requirements to D is quite tricky: how do we know whether the stationary regime has been established? how do we know which site started the infection? and what if infection has been started by a set of sites? The authors resolve this questions by employing the shape theorem (Theorem 1.2 in the paper,e.g. [5]). The meaning of this theorem is very well described in the paper right after its statement on p. 1073. In summary, the shape theorem has two consequences. First, the set of infected sites grows with time t roughly as tU where U is a non-random set. Second, inside U, the processes started with one infected site and with all infected sites are equal eventually almost surely. Both consequences are extremely important for establishing the results of the paper. In Theorem 2.1, the authors prove that for the process started with one infected site, the convex hull of all infected sites is squeezed between $(1-\varepsilon)Ut$ and $(1+\varepsilon)Ut$ eventually a.s. provided that the process survives forever. Thus, the convex hull of infected sites becomes a starting point for creating a suitable set D. Next, the similarity of the process started with one and all infected sites provides the tool for proving the consistency of the estimator as $t \to \infty$, see Section 4.

Two other important elements of the model and the approach must be mentioned: shrinking and bounded correlations. Throughout the paper authors work not directly with the convex hull of infected sites $\mathscr{C}_t$ but rather with a so-called shrinking of this set, C_t. Shrinking is defied in Section 3 in a very general sense, and several possible procedures are suggested to obtain a shrinking. Essentially, shrinking means that the 'border' sites have to be removed. The reason is that the equilibrium has not yet set on these sites, and this may (and will!) distort the estimator. The consistency of the estimator (23.1) with $D = C_t$ as $t \to \infty$ holds under very mild shrinking conditions. However, for the asymptotic normality to hold, a certain fraction of nodes from $\mathscr{C}_t$ has to be removed. The authors notice that in fact, to obtain a good estimator, one

should remove 20% to 40% of sites. Further, for the asymptotic normality it is crucial that correlations between any two sites decrease exponentially with the distance between these sites. These short-range dependencies, that ensure that some sort of central limit theorem must hold, are stated in Theorem 2.2 and further in Lemma 5.1. The asymptotic normality of the estimator is established in Theorem 5.1. This is not however the end of the story because the asymptotic variance of the estimator involves unknown parameters. In Section 6 the authors discuss this difficulty and provide a possible plausible solution.

I would like to add that a quantitative analysis of infection spread is definitely a very important topic, for example, in social and computer networks. Such networks however are usually not a grid. On the contrary, they exhibit power law degree distributions and the well known small-world phenomenon. These fascinating properties of real-life networks motivated an emergence of a new research area, devoted to the studies of complex systems, that has boosted in the last ten years. We refer to e.g. [1] for a survey of the field and its relation to statistical mechanics and interacting particle systems. The problem of infection spread in complex networks is for sure one of the key topics in this new area (see e.g. [2, 6, 10, 11]). Rigorous mathematical studies in this direction have just started. Obviously, the problem of parameter estimation for existing computer viruses and pandemics is highly relevant and offers an endless number of new mathematical challenges.

23.4 Collecting n items on a circle

Finally, my own paper [8]. This work was a continuation of my PhD thesis that I did at EURANDOM, in Eindhoven. I was lucky to have a PhD project that I could explain to anyone even without the famous back-side of an envelop. Imagine a circle and suppose that n items are distributed randomly at its circumference, which we assume to have a length 1. We start at point zero and move at a constant (unit) speed with the goal to collect all n items. We may move in one direction or turn, following any strategy we like. For instance we may choose to never change a direction, or always collect an item nearest to our current position, or pick the shortest route. The problem is to find the distribution of the travel time under different strategies. The question arises in automated storage and retrieval systems known as warehousing carousels. A circle represents a carousel that consists of a large number of shelves or drawers moving in a closed loop in either direction, and the items are locations of the products to be picked. The objective is to evaluate the rotation time, which is an important part of the response time of the system.

Clearly, if we just move in one direction, the problem is trivial: the probability to collect all items within time $t \in [0,1]$ is just t^n. However, already for the nearest-item strategy a straightforward approach results in hopelessly messy calculations, which do not lead to any meaningful outcomes. Nevertheless, the problem has an elegant solution, and the distribution of the travel time often can be written in a very simple form. The fruitful idea is to recall that the intervals between adjacent items are

uniform spacings that are distributed as i.i.d. exponential random variables, divided by their sum. Then the travel time can be written as a function of exponential random variables. In order to find the distribution of this function, the memory-less property can be used yielding surprisingly simple outcomes like in Lemma 1.1 in the paper. This way, in the papers with Ivo Adan, we derived elegant formulas for the travel time distribution under the nearest item heuristic and some other close-to-optimal strategies. For the optimal route, the problem however remained open.

It may take at most one-two minutes to guess what the optimal route on a circle should be. Clearly, it is not optimal to turn more than once. Thus, we just have to choose the shortest out of the $2n$ routes with no turn or one turn. The distribution of the optimal travel time however remains tricky even if we employ the spacings. The difficulty arises from the theoretical possibility that we may have to collect more than a half of the items before the turn. Although this scenario is all but irrelevant in practice, it has to be taken into account in the analysis, messing up the calculations. In the thesis I could not solve the problem and presented only some preliminary results on the upper bounds for the optimal route (Section 2 of the paper). Willem liked the problem from the very beginning and always believed that the distribution of the optimal route can be obtained. This paper started with obtaining the recursive equation for the optimal route (Section 3). Although the equations are not explicit, we do provide a recursion, which makes it possible to find the minimal travel time distribution for any n.

The results became much cleaner and the focus of the paper actually shifted when we turned to the asymptotic behavior as n goes to infinity. Theorem 4.2 states that in this case the difference between the shortest travel time and one complete rotation behaves as $1/(n+1)$ multiplied by the maximum between two independent random variables of the form $J = \sum_{i=1}^{\infty} (2^i - 1)^{-1} X_i$, where X_i's are independent standard exponential random variables.

Interestingly, at that time such weighted sums of exponentials attracted a lot of attention as a special case of an exponential functional of a Poisson process (see Section 6). In particular, Fabrice Guillemin, Philippe Robert and Bert Zwart encountered such functionals in the analysis of a transmission control protocols on the Internet. One intriguing and unresolved question about such random variables was their lower-tail behavior, that is, the asymptotic expression of $P(J < t)$ as $t \to 0$. To this end, only the asymptotics of $\log P(J < t)$ was known. The article by Davis and Resnick that Bert Zwart pointed to us was highly relevant but the results could not be applied directly because they were given in the form of transforms. After long calculations we arrived to the formula (5.8) that provided the exact asymptotic behavior in a closed-form. Compared to the logarithmic asymptotics, this formula contained several additional terms that were not known before. However, we were not completely satisfied because one of the factors (the function C in (5.8)) was defined by an infinite product. When plotted, this function looked liked a constant. Was it a yet another weird way to write a constant? It was tempting to prove it. We were delighted when a more detailed analysis (Proposition 5.1) revealed that our function C had an unexpected oscillating behavior involving theta-functions.

The oscillations were so small that they simply could not be seen in the plots, the analysis was needed to find them!

The explanation of why the oscillations appear seems to lie in the sort of a 'binary tree structure' of our functional J, whose coefficients are negative powers of two. Later on, Philippe Robert found that such oscillating asymptotic behavior is a typical feature of algorithms with a tree structure. For further reading I recommend his very interesting papers [9] and [12]. I think that the oscillating asymptotic behavior of algorithms is a highly compelling phenomenon, and I am very happy that our paper contributed in its study.

23.5 Acknowledgement

I would like to thank Sara van de Geer and Marten Wegkamp for creating this volume and for inviting me to contribute in it.

References

1. Réka Albert and Albert-László Barabási (2002). Statistical mechanics of complex networks.*Rev. Mod. Phys.* **74** 47–97.
2. N. Berger, C. Borgs, J.T. Chayes, and A. Saberi (2005). On the spread of viruses on the Internet. In:*Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, 301310. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 301–310.
3. Mathisca C. M. de Gunst and Willem R. van Zwet (1992). A non-Markovian model for cell population growth: speed of convergence and central limit theorem. *Stochastic Process. Appl.* **41**, 297–324.
4. Mathisca C. M. de Gunst and Willem R. van Zwet (1993). A non-Markovian model for cell population growth: tail behavior and duration of the growth process. *Ann. Appl. Probab.* **3** 1112–1144.
5. Rick Durrett. The contact process, 19741989 (1991). In: *Mathematics of Random Media*, (Blacksburg, VA, 1989), volume 27 of Lectures in Appl. Math., pages 1–18. Amer. Math. Soc., Providence, RI.
6. Rick Durrett and Paul Jung (2007). Two phase transitions for the contact process on small worlds. *Stochast. Process. Appl.* **117**, 1910–1927.
7. Marta Fiocco and Willem R. van Zwet (2003). Parameter estimation for the supercritical contact process. *Bernoulli* **9** 1071–1092.
8. N. Litvak and W. R. van Zwet (2004). On the minimal travel time needed to collect n items on a circle. *Ann. Appl. Probab.* **14** 881–902.
9. Hanène Mohamed and Philippe Robert (2005). A probabilistic analysis of some tree algorithms. *Ann. Appl. Probab.* **15** 2445–2471.
10. M. E. J. Newman. Spread of epidemic disease on networks (2002). *Physical Review E* **66** 16–28.
11. R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks (2001). *Physical review letters* **86** 3200–3203.
12. Philippe Robert. On the asymptotic behavior of some algorithms (2005). *Random Structures Algorithms* **27** 235–250.
13. Sheldon M. Ross (1996). *Stochastic processes*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition.